# Penerapan Data Science pada Analisis Data Acara TV dan Film pada Aplikasi Layanan *Streaming*

Sienie Celicia Dewi[#1], Hendra Bunyamin[*2], Setia Budi[#3]

[#] *Sistem Informasi, Universitas Kristen Maranatha*
*Jalan Surya Sumantri no. 65, Bandung, Indonesia.*
[1]1873006@maranatha.ac.id
[3]setia.budi@it.maranatha.edu

[*]*Sistem Informasi, Universitas Kristen Maranatha*
*Jalan Surya Sumantri no. 65, Bandung, Indonesia.*
[2] hendra.bunyamin@it.maranatha.edu

*Abstract* — **Data science is one of the sciences that can be applied to analyze data and explore data on a large scale. This study was conducted to analyze the data on TV shows and movies contained on Netflix. The dataset in this study was obtained from the Kaggle.com site uploaded by Shivam Bansal. The dataset has information from existing shows such as title, name of the director, actress and actor, country of origin, duration, genre, and description. This research was conducted on data from 2016 to 2020 by applying the stages of Exploratory Data Analysis which started with the stage of describing the problem until the results were obtained which were then displayed in the form of data visualization to help summarize the results. In this study, data analysis was also carried out to display recommendations for an impression using the Cosine Similarity algorithm, applying Clustering using K-Means, and N-Gram from the results of Clustering.**

*Keywords*— **Clustering, Cosine Similarity, Data Science, EDA.**

## I. PENDAHULUAN

Salah satu hal yang dapat kita temukan di Internet dalam kategori hiburan yaitu layanan untuk *streaming* film dan acara TV. Netflix adalah salah satu layanan *streaming* acara TV dan film yang berasal dari Amerika Serikat. Netflix merupakan salah satu layanan *streaming* yang menyediakan acara TV dan film dengan berbagai genre dan rating usia bagi pengguna layanan Netflix dari berbagai usia. Netflix menjadi penyedia layanan *streaming* acara TV dan film yang memiliki pelanggan hingga 204 juta pengguna [1].

Dalam penerapan dari analisis *data science*, diperlukan *dataset* yang mumpuni untuk mendapatkan hasil analisis yang maksimal. *Dataset* yang digunakan yaitu "Netflix Movies and TV Shows"[1] oleh Shivam Bansal yang diambil dari situs Kaggle. Dalam *dataset* yang digunakan dalam penelitian ini terdapat 12 informasi yang tersedia untuk dianalisis, yaitu judul, nama sutradara, aktris dan aktor, negara asal, tanggal ditambahkan ke Netflix, tanggal rilis acara TV atau film, rating usia, durasi, genre, dan deskripsi dari acara TV atau film.

Penerapan *data science* pada data Netflix memungkinkan untuk mendapatkan hasil analisis terhadap film atau acara TV yang terdapat di Netflix. Salah satu hasil yang dapat di analisis adalah data dari aktor atau aktris dari film atau acara TV pada Netflix. Hasil analisis ini dapat menunjukkan bahwa aktor dan aktris tersebut merupakan aktor dan aktris yang berkesan sehingga dia sering muncul di beberapa film ataupun acara TV. Setelah data dapat di analisis, data akan divisualisasikan ke dalam bentuk grafik untuk membantu setiap orang melihat kesimpulan dari hasil analisis.

Penelitian ini dilakukan untuk mendapatkan pengetahuan mengenai tahapan untuk melakukan *Exploratory Data Analysis*, melakukan eksplorasi terhadap data Movies dan TV Shows Netflix, dan melakukan visualisasi terhadap *dataset* yang telah dieksplorasi.

## II. KAJIAN TEORI

### *A. Data Science*

[1] www.kaggle.com/shivamb/netflix-shows

Data Science merupakan ilmu yang menggabungkan ilmu matematika dan statistika dengan menggunakan ilmu komputer untuk dapat melakukan pembelajaran, analisis, dan evaluasi suatu data. Data science menggunakan pendekatan multidisiplin untuk mendapatkan pengetahuan yang dapat direpresentasikan dari suatu data [2, 3].

Tahapan awal yang dilakukan dalam analisis pada data science yaitu EDA atau Exploratory Data Analysis. EDA merupakan sebuah pendekatan yang membantu menganalisis dataset untuk dapat meringkas data dengan cara menggunakan karakteristik dari data dan memodelkannya dalam bentuk visualisasi data dengan menggunakan penggambaran yang tepat [4]. Penggambaran hasil analisis dengan visualisasi data diharapkan dapat membantu orang yang melihat gambar dan juga menarik kesimpulan dengan tepat[5]. Beberapa contoh untuk melakukan visualisasi data yaitu menggunakan diagram batang, diagram pie, *scatter plot*, dan sebagainya.

Berikut ini adalah empat langkah yang dilakukan selama EDA[6].

1. Pendefinisian masalah
2. Persiapan data
3. Analisis data
4. Pengembangan dan representasi hasil

### *B. Pandas*

Pandas merupakan sebuah *library* berlisensi BSD dan *open source*. Pandas menyediakan struktur data dan analisis yang mudah digunakan. File yang dapat dibaca oleh Pandas diantaranya .txt, .csv, .tsv. [7]

### *C. Matplotlib*

Matplotlib merupakan salah satu *library* Python yang terkenal untuk melakukan visualisasi data yaitu membuat plot dan visualisasi dua dimensi [8].

### *D. Seaborn*

Seaborn merupakan *library* untuk visualisasi data yang dirancang di atas Matplotlib. Seaborn disebut juga sebagai *wrapper* bagi Matplotlib karena Seaborn dirancang agar dapat menggunakan manfaat dari Matplotlib, tetapi dengan cara yang lebih efisien [9].

### *E. Cosine Similarity*

Kesamaan kosinus atau *cosine similarity* adalah sebuah metode yang dapat digunakan untuk mengukur kemiripan antara dua vektor bukan nol dari hasil kali yang mengukur sudut cosinus di antara dua vektor [10].

### *F. Scikit-Learn*

Scikit-Learn merupakan *library* untuk machine learning yang mendukung proses dengan teknik *supervised* dan *unsupervised learning* yang dapat digunakan secara gratis. Scikit-Learn dilengkapi berbagai alat yang dapat digunakan untuk melakukan *fitting model*, *preprocessing*, *model selection* dan *evaluation*, *clustering*, *classification* dan sebagainya [11].

### *G. Count Vectorizer*

Count Vectorizer adalah sebuah *tools* yang disediakan oleh *library* Scikit-Learn yang dapat digunakan untuk mentransformasikan data teks ke bentuk vektor berdasarkan frekuensi dari setiap kata yang muncul dalam keseluruhan teks. Dengan menggunakan Count Vectorizer, sebuah matriks akan dibuat untuk menampung setiap kata dalam huruf kecil secara unik dan terurut secara alphabetic pada setiap kolom, teks dari dokumen ditampung dalam baris, dan setiap *cell* akan diisi dengan jumlah kemunculan kata dalam teks pada dokumen [12].

### *H. TF-IDF Vectorizer*

TF-IDF atau *Term Frequency-Inverse Document Frequency* adalah cara untuk menghitung kuantitas suatu kata dalam sekumpulan dokumen. Setiap kata akan dihitung skor yang menunjukkan pentingnya suatu kata dalam dokumen. TF-IDF digunakan untuk mentransformasikan kata menjadi bentuk numerik[13].

### *I. NLTK*

NLTK atau Natural Language Toolkit adalah platform pada Python yang digunakan untuk melakukan proses pada data dengan jenis teks. NLTK memiliki berbagai library yang dapat digunakan, misalnya untuk text processing seperti klasifikasi, tokenisasi, stemming, tagging dan sebagainya. NLTK dapat digunakan secara gratis. Untuk menggunakan NLTK, instalasi dapat dilakukan dengan tahapan yang terdapat pada dokumentasi dari NLTK [14].

### *J. K-Means Clustering*

Clustering adalah proses untuk melakukan pengelompokkan terhadap suatu data berdasarkan kemiripan setiap data. Umumnya clustering digunakan untuk mengeksplorasi dataset untuk melihat pola ataupun karakteristik [15]. K-Means adalah algoritma yang digunakan untuk melakukan *clustering*, dalam K-Means jumlah cluster diwakili oleh variabel K. Dalam prosesnya, algoritma K-Means melakukan iterasi terhadap penempatan nilai pusat atau yang dikenal dengan cluster center, cluster akan terbentuk ketika proses tersebut selesai [15, 16].

### *K. N-Gram*

N-Gram adalah *modeling language* yang memberikan probabilitas untuk kalimat dan urutan kata. N-Gram banyak digunakan dalam text mining dan Natural Language Processing [17]. N-Gram merupakan urutan dari n kata dimana n adalah jumlah urutan dari 1 hingga tak terhingga. Untuk n=1 disebut Unigram yang terdiri dari 1 kata, untuk n=2 disebut Bigram yang terdiri dari 2 kombinasi kata, n=3 disebut Trigram yang terdiri dari kombinasi 3 kata, dan seterusnya [18].

## III. METODOLOGI

### *A. Dataset*

*Dataset* "Netflix Movies and TV Shows" berisi daftar acara TV dan film dari berbagai negara di dunia seperti United States, India, United Kingdom, Japan, South Korea, Canada, Spain, France, Egypt, dan Mexico. *Dataset* ini memiliki dimensi 7787 baris dan 12 kolom. Jupyter Notebook digunakan dalam penelitian ini karena Jupyter Notebook cocok untuk melakukan analisis data dengan menggunakan bahasa pemrograman Python.

### *B. Exploratory Data Analysis (EDA)*

Untuk memulai EDA, perlu dilakukan *import library* yang dibutuhkan dalam eksplorasi. *Library* yang diimpor adalah Pandas untuk pengolahan *dataset*. *Library* akan ditambahkan sesuai selama eksplorasi berlangsung sesuai kebutuhan. Setelah melakukan *import library*, *dataset* dari format .csv diimpor ke dalam Jupyter Notebook. Setelah berhasil diimpor untuk mempersiapkan data, dilakukan pembersihan data untuk menghapus nilai *Null* dan penyesuaian tipe data. Karena dataset yang dimiliki berisi keseluruhan data hingga 2021 maka dilakukan pembatasan data menyesuaikan dengan batasan dan tujuan penelitian yaitu tahun 2016 hingga 2020.

*1) Eksplorasi berdasarkan Keseluruhan Tahun:* Eksplorasi dilakukan untuk mengetahui jumlah data yang ada di setiap tahunnya, penghitungan dilakukan dengan memanfaatkan library Pandas. Kemudian, hasil hitungnya dituangkan dalam visualisasi dengan menambahkan *library* untuk visualisasi yaitu dapat menggunakan Matplotlib atau Seaborn untuk melihat dalam ringkasan diagram.

*2) Eksplorasi berdasarkan Negara:* Eksplorasi dilakukan untuk mengetahui 5 negara dengan jumlah acara TV dan film terbanyak berdasarkan keseluruhan tahun dengan memanfaatkan *library* yang telah diimpor. Hasilnya divisualisasikan dalam bentuk diagram.

*3) Eksplorasi berdasarkan Rating Usia:* Eksplorasi dilakukan untuk mengetahui jumlah acara TV dan film berdasarkan *rating* usia berdasarkan keseluruhan tahun dengan memanfaatkan *library* yang telah diimpor. Hasilnya divisualisasikan dalam bentuk diagram.

*4) Eksplorasi berdasarkan Genre:* Eksplorasi dilakukan untuk mengetahui *genre* yang populer dari berdasarkan keseluruhan tahun dengan memanfaatkan *library* yang telah diimpor. Hasilnya divisualisasikan dalam tampilan WordCloud.

### *C. Sistem Rekomendasi*

Pada bagian ini, setelah melakukan EDA pada eksplorasi awal, dilakukan eksplorasi untuk membuat sistem rekomendasi dengan menggunakan algoritma Cosine Similarity dari *library* Scikit-Learn. Fitur-fitur yang akan dipilih sebagai acuan rekomendasi akan diubah ke dalam bentuk vektor dengan menggunakan CountVectorizer dan TF-IDF Vectorizer dengan mengimpor dari library Scikit-Learn.

### *D. Clustering dan N-Gram*

Pada bagian ini, akan dilakukan *clustering* terhadap data berdasarkan dari deskripsi setiap tayangan, dalam proses *clustering* ini digunakan library dari NLTK untuk melakukan persiapan terhadap kalimat dalam data deskripsi atau bisa disebut *preprocessing*. Hal ini perlu dilakukan karena untuk mendapatkan hasil yang maksimal, data teks harus bersih tanpa simbol, *stopword* yaitu dalam bahasa Inggris seperti *an*, *a*, *the* dan sebagainya, dan merupakan kata dasar. Kemudian, untuk

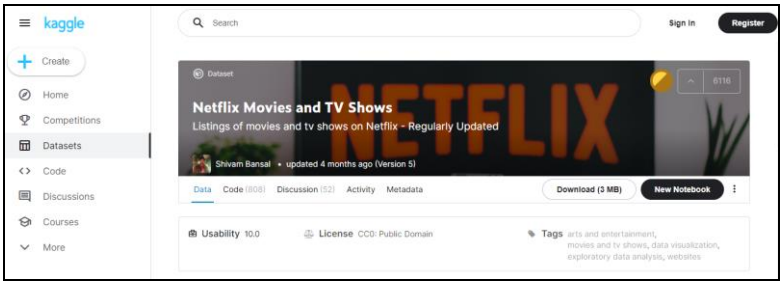
proses *clustering* digunakan algoritma K-Means dari *library* Scikit-Learn. Setelah dibagi ke dalam masing-masing *cluster*, pada cluster dengan data terbanyak akan dilakukan proses pembuatan N-Gram menggunakan library dari NLTK. N-Gram yang akan dibuat pada eksplorasi ini yaitu Bigrams dan Trigrams.

## IV. HASIL IMPLEMENTASI

### A. *Dataset*

Pada penelitian ini, *dataset* diambil dari situs Kaggle, dengan judul *dataset* "Netflix Movies dan TV Shows", *dataset* ini terus ter*update* dan saat penelitian ini dimulai jumlah baris *dataset* adalah 7787 baris dan 12 kolom. Berikut ini penjelasan dari setiap kolomnya.

1. show_id (ID dari acara TV dan film) : object
2. type (tipe dari tayangan, yaitu acara TV atau film) : object
3. title (judul dari acara TV atau film) : object
4. director (sutradara dari acara TV atau film) : object
5. cast (aktor dan aktris dalam acara TV atau film) : object
6. country (asal negara dari acara TV atau film) : object
7. date_added (tanggal acara TV atau film ditambahkan ke Netflix) : object
8. release_year (tahun acara TV atau film dirilis) : int
9. rating (rating usia dari acara TV atau film): object
10. duration (durasi dari acara TV atau film) : object
11. listed_in (genre dari acara TV atau film) : object
12. description (deskripsi singkat dari acara TV atau film) : object




Gambar 1. Sumber *Dataset*

Berikut ini adalah tampilan 5 data pertama hasil dari *load dataset*, Gambar 2 merupakan hasil dari *load dataset* yang ditampilkan dalam bentuk Dataframe dari Pandas dengan jumlah kolom 12.

| | show_id | type | title | director | cast | country | date_added | release_year | rating | duration | listed_in | description |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | s1 | TV Show | 3% | NaN | João Miguel, Bianca Comparato, Michel Gomes, R... | Brazil | August 14, 2020 | 2020 | TV-MA | 4 Seasons | International TV Shows, TV Dramas, TV Sci-Fi &... | In a future where the elite inhabit an island ... |
| 1 | s2 | Movie | 7:19 | Jorge Michel Grau | Demián Bichir, Héctor Bonilla, Oscar Serrano, ... | Mexico | December 23, 2016 | 2016 | TV-MA | 93 min | Dramas, International Movies | After a devastating earthquake hits Mexico Cit... |
| 2 | s3 | Movie | 23:59 | Gilbert Chan | Tedd Chan, Stella Chung, Henley Hii, Lawrence ... | Singapore | December 20, 2018 | 2011 | R | 78 min | Horror Movies, International Movies | When an army recruit is found dead, his fellow... |
| 3 | s4 | Movie | 9 | Shane Acker | Elijah Wood, John C. Reilly, Jennifer Connelly... | United States | November 16, 2017 | 2009 | PG-13 | 80 min | Action & Adventure, Independent Movies, Sci-Fi... | In a postapocalyptic world, rag-doll robots hi... |
| 4 | s5 | Movie | 21 | Robert Luketic | Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar... | United States | January 1, 2020 | 2008 | PG-13 | 123 min | Dramas | A brilliant group of students become card-coun... |

Gambar 2. *DataFrame*

### B. *Exploratory Data Analysis (EDA)*

Untuk mengetahui karakteristik data, dapat menggunakan kode `.dtypes` untuk menampilkan tipe data dari setiap kolom, Tipe data dari setiap kolom pada dataset ini dapat dilihat pada bagian IV.A. Setelah mengetahui setiap tipe data pada kolom, dilakukan pembersihan data, pembatasan data, dan perubahan tipe data untuk menyesuaikan dengan datanya.

Setelah dilakukan pembersihan dan pembatasan data, jumlah baris menjadi 4639 dan 12 kolom, berikut ini pada Gambar 3 merupakan hasil dimensi setelah dilakukan pembersihan dan pembatasan data.

```
fiveyearadded['date_added'].unique()
```

```
array([2016, 2018, 2017, 2020, 2019], dtype=int64)
```

```
fiveyearadded.shape
```

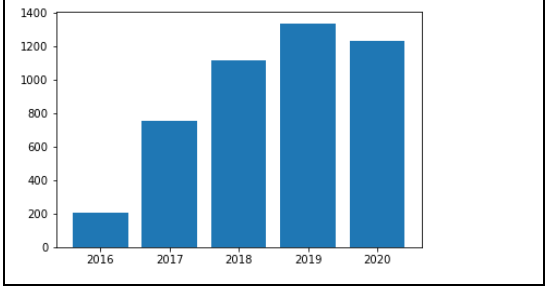
```
(4639, 12)
```

Gambar 3. Dimensi Dataset

Kemudian setelah melakukan pembersihan data dan pembatasan data, dilakukan penyesuaian tipe data. Sebelumnya dari 12 kolom terdapat 1 kolom yang memiliki tipe data `int`, yaitu kolom `release_year` dan untuk 11 kolom lainnya yaitu `show_id`, `type`, `title`, `director`, `cast`, `country`, `date_added`, `rating`, `duration`, `listed_in` dan `descrption` memiliki tipe data `object`. Setelah dilakukan penyesuaian tipe data, dilakukan perubahan tipe data untuk kolom `date_added` dan `release_year` menjadi tipe data `datetime`, untuk kolom lainnya tetap `object`.

*1) Eksplorasi berdasarkan Keseluruhan Tahun:* Eksplorasi pada bagian ini dilakukan untuk melihat jumlah data dari tahun 2016 hingga 2020, Gambar 4 merupakan hasil dari jumlah data setiap tahunnya. Gambar tersebut menunjukkan jumlah dari tayangan di setiap tahunnya dari tahun 2016, 2017, 2018, 2019 dan 2020 dan dari hasil tersebut didapatkan data cenderung naik dan mengalami sedikit penurunan di tahun 2020.

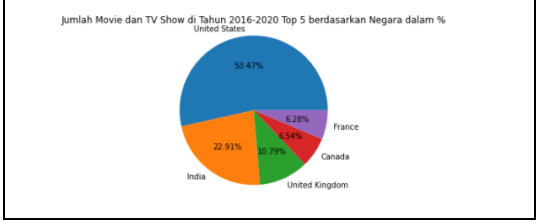
```
totaltahun = fiveyearadded.groupby(['date_added']).count()
totaltahun = totaltahun.filter(['show_id'])
totaltahun.columns = ['count']
pd.DataFrame(totaltahun)
plt.bar(totaltahun.index, totaltahun['count'])
plt.show()
```




Gambar 4. Jumlah Tayangan berdasarkan Keseluruhan Tahun

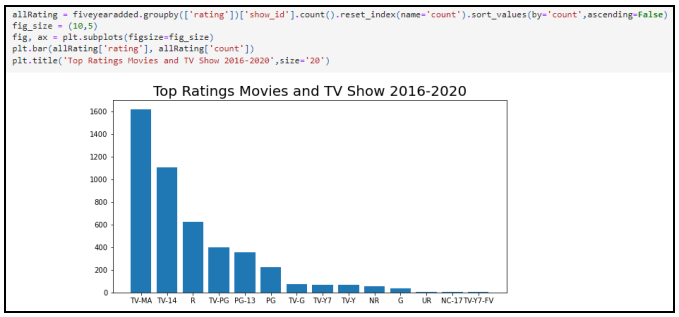
*2) Eksplorasi berdasarkan Negara*: Berikut ini adalah hasil dari eksplorasi terhadap 5 negara dengan jumlah tayangan terbanyak pada keseluruhan tahun, Gambar 5 merupakan diagram pie yang menunjukkan 5 negara dengan jumlah tayangan terbanyak. Bagian yang berwarna biru dengan label United States menunjukkan bahwa dalam persentase, kontribusi negara Unites States sebesar 53,47%, bagian berwarna kuning untuk negara India dengan persentase 22,91%, bagian berwarna hijau untuk negara United Kingdom dengan persentase 10.79%, bagian berwarna merah untuk negara Canada dengan persentase 6,54%, dan bagian berwarna ungu untuk negara France dengan persentase 6,28%. Dalam pie chart, semakin besar bagian yang dimiliki suatu label menandakan semakin dominan juga data yang direpresentasikan dengan label tersebut.

```
persentase0 = countNegara[1].head()
label= countNegara[0].head()
plt.title('Jumlah Movie dan TV Show di Tahun 2016-2020 Top 5 berdasarkan Negara dalam %')
plt.pie((persentase0), labels=label, autopct='%.2f%%')
plt.axis('equal')
plt.show()
```




Gambar 5. 5 Negara dengan Jumlah Tayangan Terbanyak berdasarkan Keseluruhan Tahun

*3) Eksplorasi berdasarkan Rating Usia*: Berikut ini adalah hasil dari eksplorasi terhadap jumlah tayangan berdasarkan rating usia. Gambar 6 merupakan merupakan diagram batang yang menunjukkan tayangan berdasarkan rating usia. Sumbu x merepresentasikan rating dari tayangan dan sumbu y merepresentasikan jumlah tayangan yang ada. Hasil visualisasi ini dibuat secara terurut dari rating dengan tayangan paling banyak hingga rating dengan tayangan paling sedikit. Dari hasil visualisasi ini dapat dilihat untuk 5 rating dengan tayangan terbanyak dari keseluruhan tahun, yaitu rating TV-MA, TV-14, R, TV-PG, dan PG-13




Gambar 6. Hasil Tayangan berdasarkan Rating Usia

*4) Eksplorasi berdasarkan Genre*: Berikut ini adalah hasil dari eksplorasi terhadap jumlah tayangan berdasarkan *genre*. Gambar 7 merupakan visualisasi dalam bentuk WordCloud yang menunjukkan genre yang terdapat pada keseluruhan data, pada WordCloud kata-kata yang paling besar ukurannya merupakan kata-kata yang paling dominan pada data.


Gambar 7. WordCloud Genre

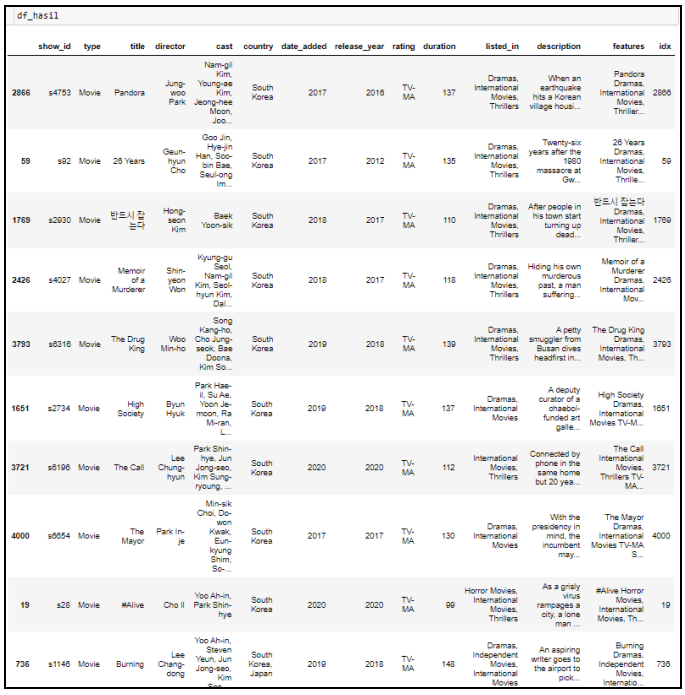
Dalam penggunaan WordCloud dengan data yang jumlah selisihnya tidak terlalu jauh mengakibatkan perbandingan ukuran tulisan dalam WordCloud tidak terlalu terlihat. Maka dari itu pada Gambar 8 merupakan hasil hitung yang menunjukkan 5 *genre* teratas pada data. *Genre* dengan tayangan terbanyak dimiliki oleh *genre* International Movie dengan jumlah 2151.

```
[('International Movies', 2151),
 ('Dramas', 1977),
 ('Comedies', 1353),
 ('Action & Adventure', 678),
 ('Independent Movies', 650),
```

Gambar 8. Hasil Hitung *Genre* Keseluruhan Tahun

## *C. Sistem Rekomendasi*

Setelah melakukan eksplorasi awal terhadap data, berikut ini merupakan hasil pembuatan sistem rekomendasi dengan menggunakan algoritma Cosine Similarity. Fitur yang digunakan sebagai acuan dalam pembuatan rekomendasi ini adalah kolom judul, genre, rating usia dan negara. Berikut ini pada Gambar 9 merupakan hasil 10 rekomendasi terhadap tayangan dengan judul Forgotten dengan menggunakan vektorisasi CountVectorizer. Secara terurut, tayangan paling atas pada gambar tersebut adalah rekomendasi terdekat dengan skor terbesar untuk tayangan dengan judul Forgotten.

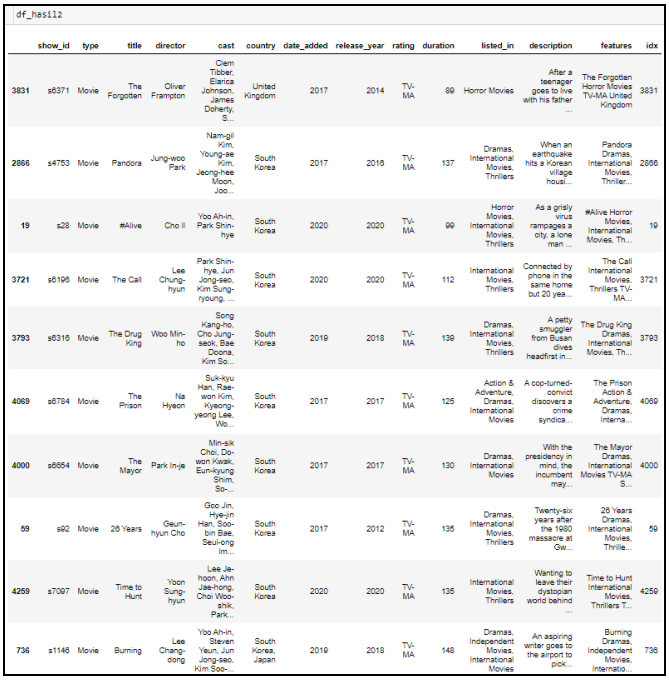


df_hasil

| | show_id | type | title | director | cast | country | date_added | release_year | rating | duration | listed_in | description | features | idx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2866 | s4753 | Movie | Pandora | Jung-woo Park | Nam-gil Kim, Young-ae Kim, Jeong-hee Moon, Joo... | South Korea | 2017 | 2016 | TV-MA | 137 | Dramas, International Movies, Thrillers | When an earthquake hits a Korean village housi... | Pandora Dramas, International Movies, Thriller... | 2866 |
| 59 | s92 | Movie | 26 Years | Geun-hyun Cho | Goo Jin, Hye-jin Han, Soo-bin Bae, Seul-ong Im... | South Korea | 2017 | 2012 | TV-MA | 135 | Dramas, International Movies, Thrillers | Twenty-six years after the 1980 massacre at Gw... | 26 Years Dramas, International Movies, Thrille... | 59 |
| 1769 | s2930 | Movie | 반드시 잡는다 | Hong-seon Kim | Baek Yoon-sik | South Korea | 2018 | 2017 | TV-MA | 110 | Dramas, International Movies, Thrillers | After people in his town start turning up dead... | 반드시 잡는다 Dramas, International Movies, Thriller... | 1769 |
| 2426 | s4027 | Movie | Memoir of a Murderer | Shin-yeon Won | Kyung-gu Seol, Nam-gil Kim, Seol-hyun Kim, Dal... | South Korea | 2018 | 2017 | TV-MA | 118 | Dramas, International Movies, Thrillers | Hiding his own murderous past, a man suffering... | Memoir of a Murderer Dramas, International Mov... | 2426 |
| 3793 | s6316 | Movie | The Drug King | Woo Min-ho | Song Kang-ho, Cho Jung-seok, Bae Doona, Kim So... | South Korea | 2019 | 2018 | TV-MA | 139 | Dramas, International Movies, Thrillers | A petty smuggler from Busan dives headfirst in... | The Drug King Dramas, International Movies, Th... | 3793 |
| 1651 | s2734 | Movie | High Society | Byun Hyuk | Park Hae-il, Su Ae, Yoon Je-moon, Ra Mi-ran, L... | South Korea | 2019 | 2018 | TV-MA | 137 | Dramas, International Movies | A deputy curator of a chaebol-funded art galle... | High Society Dramas, International Movies TV-M... | 1651 |
| 3721 | s6196 | Movie | The Call | Lee Chung-hyun | Park Shin-hye, Jun Jong-seo, Kim Sung-ryoung, ... | South Korea | 2020 | 2020 | TV-MA | 112 | International Movies, Thrillers | Connected by phone in the same home but 20 yea... | The Call International Movies, Thrillers TV-MA... | 3721 |
| 4000 | s6654 | Movie | The Mayor | Park In-je | Min-sik Choi, Do-won Kwak, Eun-kyung Shim, So-... | South Korea | 2017 | 2017 | TV-MA | 130 | Dramas, International Movies | With the presidency in mind, the incumbent may... | The Mayor Dramas, International Movies TV-MA S... | 4000 |
| 19 | s28 | Movie | #Alive | Cho Il | Yoo Ah-in, Park Shin-hye | South Korea | 2020 | 2020 | TV-MA | 99 | Horror Movies, International Movies, Thrillers | As a grisly virus rampages a city, a lone man ... | #Alive Horror Movies, International Movies, Th... | 19 |
| 736 | s1146 | Movie | Burning | Lee Chang-dong | Yoo Ah-in, Steven Yeun, Jun Jong-seo, Kim So... | South Korea, Japan | 2019 | 2018 | TV-MA | 148 | Dramas, Independent Movies, International Movies | An aspiring writer goes to the airport to pick... | Burning Dramas, Independent Movies, Internatio... | 736 |

Gambar 9. Rekomendasi CountVectorizer

Berikutnya pada Gambar 10 merupakan hasil 10 rekomendasi terhadap tayangan dengan judul Forgotten dengan menggunakan vektorisasi TF-IDF Vectorizer, secara terurut juga tayangan paling atas pada gambar tersebut merupakan rekomendasi terdekat dengan skor terbesar untuk tayangan dengan judul Forgotten.



df_hasil2

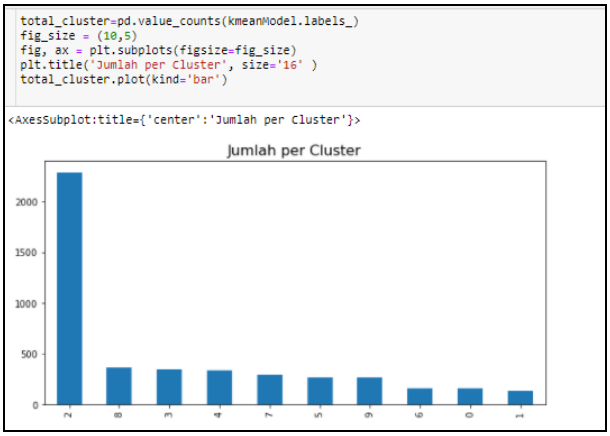
| | show_id | type | title | director | cast | country | date_added | release_year | rating | duration | listed_in | description | features | idx |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3831 | s6371 | Movie | The Forgotten | Oliver Frampton | Clem Tibber, Elarica Johnson, James Doherty, S... | United Kingdom | 2017 | 2014 | TV-MA | 89 | Horror Movies | After a teenager goes to live with his father ... | The Forgotten Horror Movies TV-MA United Kingdom | 3831 |
| 2866 | s4753 | Movie | Pandora | Jung-woo Park | Nam-gil Kim, Young-ae Kim, Jeong-hee Moon, Joo... | South Korea | 2017 | 2016 | TV-MA | 137 | Dramas, International Movies, Thrillers | When an earthquake hits a Korean village housi... | Pandora Dramas, International Movies, Thriller... | 2866 |
| 19 | s28 | Movie | #Alive | Cho Il | Yoo Ah-in, Park Shin-hye | South Korea | 2020 | 2020 | TV-MA | 99 | Horror Movies, International Movies, Thrillers | As a grisly virus rampages a city, a lone man ... | #Alive Horror Movies, International Movies, Th... | 19 |
| 3721 | s6196 | Movie | The Call | Lee Chung-hyun | Park Shin-hye, Jun Jong-seo, Kim Sung-ryoung, ... | South Korea | 2020 | 2020 | TV-MA | 112 | International Movies, Thrillers | Connected by phone in the same home but 20 yea... | The Call International Movies, Thrillers TV-MA... | 3721 |
| 3793 | s6316 | Movie | The Drug King | Woo Min-ho | Song Kang-ho, Cho Jung-seok, Bae Doona, Kim So... | South Korea | 2019 | 2018 | TV-MA | 139 | Dramas, International Movies, Thrillers | A petty smuggler from Busan dives headfirst in... | The Drug King Dramas, International Movies, Th... | 3793 |
| 4069 | s6784 | Movie | The Prison | Na Hyeon | Suk-kyu Han, Rae-won Kim, Kyeong-yeong Lee, Wo... | South Korea | 2017 | 2017 | TV-MA | 125 | Action & Adventure, Dramas, International Movies | A cop-turned-convict discovers a crime syndica... | The Prison Action & Adventure, Dramas, Interna... | 4069 |
| 4000 | s6654 | Movie | The Mayor | Park In-je | Min-sik Choi, Do-won Kwak, Eun-kyung Shim, So-... | South Korea | 2017 | 2017 | TV-MA | 130 | Dramas, International Movies | With the presidency in mind, the incumbent may... | The Mayor Dramas, International Movies TV-MA S... | 4000 |
| 59 | s92 | Movie | 26 Years | Geun-hyun Cho | Goo Jin, Hye-jin Han, Soo-bin Bae, Seul-ong Im... | South Korea | 2017 | 2012 | TV-MA | 135 | Dramas, International Movies, Thrillers | Twenty-six years after the 1980 massacre at Gw... | 26 Years Dramas, International Movies, Thrille... | 59 |
| 4259 | s7097 | Movie | Time to Hunt | Yoon Sung-hyun | Lee Je-hoon, Ahn Jae-hong, Choi Woo-shik, Park... | South Korea | 2020 | 2020 | TV-MA | 135 | International Movies, Thrillers | Wanting to leave their dystopian world behind ... | Time to Hunt International Movies, Thrillers T... | 4259 |
| 736 | s1146 | Movie | Burning | Lee Chang-dong | Yoo Ah-in, Steven Yeun, Jun Jong-seo, Kim Soo-... | South Korea, Japan | 2019 | 2018 | TV-MA | 148 | Dramas, Independent Movies, International Movies | An aspiring writer goes to the airport to pick ... | Burning Dramas, Independent Movies, Internatio... | 736 |

Gambar 10. Rekomendasi TF-IDF Vectorizer

Dari hasil kedua cara vektorisasi ini didapatkan hasil dari rekomendasi yang kurang lebih mirip walaupun menggunakan cara vektorisasi yang berbeda.

*D. Clustering dan N-Gram*

Pada bagian ini akan ditampilkan hasil pembagian cluster beserta jumlahnya disetiap cluster. Pada eksplorasi ini data dibagi ke 10 cluster, Gambar 11 merupakan diagram batang yang merepresentasikan jumlah di setiap clusternya. Berdasarkan gambar tersebut, cluster 2 memiliki data paling banyak dibandingkan dengan cluster lain.




Gambar 11. Hasil Clustering

Karena cluster 2 memiliki data paling banyak, berikut ini pada Gambar 12 adalah hasil Bigrams dan Trigrams untuk cluster 2 secara terurut dari yang paling banyak muncul.

```
bigrams2 = nltk.bigrams(textng2)
freqb2 = nltk.FreqDist(bigrams2)
freqb2.most_common(5)
```

```
[(('base', 'true'), 16),
 (('fall', 'love'), 16),
 (('join', 'force'), 13),
 (('take', 'stage'), 13),
 (('world', 'war'), 12)]
```

```
trigrams2 = nltk.trigrams(textng2)
freqt2 = nltk.FreqDist(trigrams2)
freqt2.most_common(5)
```

```
[(('world', 'war', 'ii'), 6),
 (('must', 'decide', 'whether'), 4),
 (('travel', 'back', 'time'), 3),
 (('take', 'stage', 'brooklyn'), 3),
 (('game', 'cat', 'mouse'), 3)]
```

Gambar 12. Hasil N-Gram Cluster 2

## V. KESIMPULAN

Setelah melakukan eksplorasi terhadap data Movies dan TV Shows Netflix, kesimpulan yang didapatkan, yaitu tahapan Exploratory Data Analysis dilakukan dimulai dari penentuan masalah, persiapan *dataset* termasuk didalamnya yaitu penghapusan nilai Null, penghapusan stopwords, dan perubahan bentuk kata, hingga pada akhirnya *dataset* siap untuk digunakan untuk proses analisis dan dari hasil proses tersebut dapat direpresentasikan ke dalam visualisasi. Berbagai informasi didapatkan dari *dataset* ini dengan memanfaatkan *library* yang tersedia seperti Pandas, Scikit-Learn, dan sebagainya, untuk menggali informasi dan dianalisis, didukung juga dengan *library* untuk visualisasi seperti Matplotlib dan Seaborn. Berbagai pilihan untuk menampilkan hasil analisis dalam visualisasi seperti diagram batang dan diagram pie.

## DAFTAR PUSTAKA

[1] F. Yolandha, "Per Akhir 2020, Jumlah Pelanggan Netflix Capai 200 Juta," Republika, 20 Januari 2020. [Online]. Available: https://www.republika.co.id/berita/qn8arz370/per-akhir-2020-jumlah-pelanggan-netflix-capai-200-juta. [Diakses 23 Februari 2021].
[2] IBM Cloud Education, "What is Data Science?," IBM, 15 Mei 2020. [Online]. Available: https://www.ibm.com/cloud/learn/data-science-introduction. [Diakses 22 Februari 2021].
[3] B. Bose, "What is Data Science: Meaning and Scope," Digital Vidya, 16 Mei 2018. [Online]. Available: https://www.digitalvidya.com/blog/what-is-data-science/. [Diakses 22 Februari 2021].
[4] K. Sahoo, A. K. Samal, J. Pramanik dan S. K. Pani, "Exploratory Data Analysis using Python," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, pp. 4727-4735, 2019.
[5] C. O. Wilke, Fundamentals of Data Visualization, Sebastopol: O'Reilly Media, Inc., 2019.
[6] S. K. Mukhiya dan U. Ahmed, Hands-On Exploratory Data Analysis with Python, Birmingham: Packt Publishing Ltd., 2020.
[7] I. Mutmainnah, "Mengenal Pandas Dalam Python," Medium, 6 Januari 2019. [Online]. Available: https://medium.com/@16611092/mengenal-pandas-dalam-python-cc66d0c5ea40. [Diakses 5 Maret 2021].
[8] W. McKinney, Python for Data Analysis, Sebastopol: O'Reilly Media, Inc., 2017.
[9] Skill Plus, "Pengenalan Seaborn," Skill Plus, 30 November 2020. [Online]. Available: https://skillplus.web.id/pengenalan-seaborn/#:~:text=Seaborn%20adalah%20library%20yang%20dibangun,dikatakan%20sebagai%20wrapper%20untuk%20matplotlib.. [Diakses 15 Maret 2021].
[10] P. Dangeti, Statistics for Machine Learning, Birmingham: Packt Publishing Ltd., 2017.
[11] Scikit-Learn, "Getting Started," Scikit-Learn, [Online]. Available: https://scikit-learn.org/stable/getting_started.html. [Diakses 28 Oktober 2021].
[12] K. Verma, "Using CountVectorizer to Extracting Features from Text," GeeksforGeeks, 17 Juli 2020. [Online]. Available: https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/. [Diakses 28 Oktober 2021].
[13] W. Scott, "TF-IDF from scratch in python on a real-world dataset.," Towards Data Science, 15 Februari 2019. [Online]. Available: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089. [Diakses 28 Oktober 2021].
[14] NLTK, "Documentation," NLTK, 19 Oktober 2021. [Online]. Available: nltk.org. [Diakses 28 Oktober 2021].
[15] P. Dangeti, Statistics for Machine Learning, Birmingham: Packt Publishing Ltd, 2017.
[16] R. D. Ramadhani, "Memahami K-Mean Clustering Pada Machine Learning Dengan Phyton," Medium, 6 Januari 2019. [Online]. Available: https://medium.com/@16611129/memahami-k-mean-clustering-pada-machine-learning-dengan-phyton-430323d80868. [Diakses 28 Oktober 2021].
[17] S. Kapadia, "Language Models: N-Gram," Medium, 26 Maret 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-language-models-n-gram-e323081503d9. [Diakses 28 Oktober 2021].
[18] E. C. D. Dios, "From DataFrame to N-Grams," Medium, 23 Mei 2020. [Online]. Available: https://towardsdatascience.com/from-dataframe-to-n-grams-e34e29df3460. [Diakses 28 Oktober 2021].